Exercise 2

Juuso Korhonen - 652377
ELEC-E8125 - Reinforcement Learning

September 21, 2022

1 Task 1

The estimated state values and policy after 100 iterations can be found in value_policy_100.pkl file.

1.1 Question 1.1

Q: What is the agent and the environment in this sailor gridworld?

A: Agent is the sailor (or the ship), and environment is the sea.

1.2 Question 1.2

Q: What is the state value of the harbour and rock states? Why?

A: State value of both is 0. It is because the simulation terminates when either one of the states is reached (no actions defined starting from these states).

1.3 Question 1.3

Q: Which path did the sailor choose, the safe path below the rocks, or the dangerous path between the rocks? If you change the reward for hitting the rocks to -10 (that is, make the sailor value life more), does he still choose the same path?

A: The sailor chooses the safe path below the rocks.

2 Task 2

Q: What happens if you run the algorithm for 30 iterations? Do the value function and policy still converge? Generally, which of them - the policy or value function - needs less iterations to converge, if any? Justify your answer.

A: No, the value function converges at 31 iterations. However, optimal policy has been already found at 30 (the sailor chooses the same route after 30 iterations). Overall, optimal policy converges (generally) earlier, because value function might still change even if it does not affect policy anymore.

3 Task 3

Required number of iterations for the value function to convergence is 31.

4 Task 4

Q: Evaluate your learned policy for N = 1000 episodes, and compute the discounted return of the initial state for each episode. The reward for crashing into rocks must be kept at -2 for this exercise. Report the average and standard deviation of the initial state's discounted return over the N=1000 episodes.

A: Average: 0.6934529247562079, Standard deviation: 1.3620300324414032

4.1 Question 4.1

Q: What is the relationship between the discounted return and the value function? Explain briefly.

A: discounted return is a random variable and value function is the expectation over it.

4.2 Question 4.2

Q: Imagine a reinforcement learning problem involving a robot exploring an unknown environment. Could the value iteration approach used in this exercise be applied directly to that problem? Why/why not? Which assumptions are unrealistic, if any?

A: No, because in this exercise we assumed that the environment is known (transition probabilities and rewards). This would obviously not hold in unknown environment.